日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出願年月日
Date of Application: ２０００年　７月１４日

出願番号
Application Number: 特願２０００－２１４２３８

出願人
Applicant(s): ソニー株式会社

CERTIFIED COPY OF
PRIORITY DOCUMENT

２００１年　５月１８日

特許庁長官
Commissioner,
Japan Patent Office

及川耕造

【書類名】　特許願
【整理番号】　0000551703
【提出日】　平成12年 7月14日
【あて先】　特許庁長官殿
【国際特許分類】　H04N　5/76
【発明者】
　【住所又は居所】　東京都品川区北品川６丁目７番３５号　ソニー株式会社内
　【氏名】　柴田　浩正
【発明者】
　【住所又は居所】　東京都品川区北品川６丁目７番３５号　ソニー株式会社内
　【氏名】　トビー　ウォーカー
【特許出願人】
　【識別番号】　000002185
　【氏名又は名称】　ソニー株式会社
　【代表者】　出井　伸之
【代理人】
　【識別番号】　100082131
　【弁理士】
　【氏名又は名称】　稲本　義雄
　【電話番号】　03-3369-6479
【手数料の表示】
　【予納台帳番号】　032089
　【納付金額】　21,000円
【提出物件の目録】
　【物件名】　明細書　1
　【物件名】　図面　1
　【物件名】　要約書　1

【包括委任状番号】　9708842

【プルーフの要否】　　要

【書類名】　　明細書

【発明の名称】　ＡＶ信号処理装置および方法、並びに記録媒体

【特許請求の範囲】

【請求項１】　供給されたＡＶ信号の内容の意味構造を反映するパターンを検出して解析し、意味のある区切りであるシーンを検出するＡＶ信号処理装置において、

前記ＡＶ信号を構成する一連のフレームによって形成されるセグメントの特徴量を抽出する特徴量抽出手段と、

基準となるセグメントと他のセグメントとの前記特徴量の類似性を測定するための測定基準を算出する算出手段と、

前記測定基準を用いて、前記基準となるセグメントと前記他のセグメントとの前記類似性を測定する類似性測定手段と、

類似性測定手段が測定した前記類似性を用いて、前記基準となるセグメントが前記シーンの境界である可能性を示す測定値を計算する測定値計算手段と、

前記測定値計算手段が計算した前記測定値の時間的パターンの変化を解析し、解析結果に基づいて前記基準となるセグメントが前記シーンの境界であるか否かを判定する境界判定手段と

を含むことを特徴とするＡＶ信号処理装置。

【請求項２】　前記ＡＶ信号は、映像信号および音声信号のうちの少なくとも一方を含む

ことを特徴とする請求項１に記載のＡＶ信号処理装置。

【請求項３】　前記基準となるセグメントに対応する前記測定値の変化の程度を示す強度値を計算する強度値計算手段を

さらに含むことを特徴とするＡＶ信号処理装置。

【請求項４】　前記測定値計算手段は、前記基準となるセグメントに対して、所定の時間領域内における類似セグメントを求め、前記類似セグメントの時間分布を解析し、過去と未来に存在する比率を数値化して前記測定値を計算する

ことを特徴とする請求１に記載のＡＶ信号処理装置。

【請求項５】　前記境界判定手段は、前記測定値の絶対値の総和にも基づき

、前記基準となるセグメントが前記シーンの境界であるか否かを判定する

ことを特徴とする請求１に記載のＡＶ信号処理装置。

【請求項６】　前記ＡＶ信号に映像信号が含まれる場合、映像セグメントの基本単位となるショットを検出して、前記音声セグメントを生成する音声セグメント生成手段を

さらに含むことを特徴とする請求２に記載のＡＶ信号処理装置。

【請求項７】　前記ＡＶ信号に音声信号が含まれる場合、前記音声信号の前記特徴量および無音区間のうちの少なくとも一方を用いて、音声セグメントを生成する音声セグメント生成手段を

さらに含むことを特徴とする請求２に記載のＡＶ信号処理装置。

【請求項８】　前記映像信号の前記特徴量には、少なくともカラーヒストグラムが含まれる

ことを特徴とする請求項２に記載のＡＶ信号処理装置。

【請求項９】　前記音声信号の前記特徴量には、音量およびスペクトラムのうちの少なくとも一方が含まれる

ことを特徴とする請求項２に記載のＡＶ信号処理装置。

【請求項１０】　前記境界判定手段は、予め設定され閾値と前記測定値を比較することにより、前記基準となるセグメントが前記シーンの境界であるか否かを判定する

ことを特徴とする請求１に記載のＡＶ信号処理装置。

【請求項１１】　供給されたＡＶ信号の内容の意味構造を反映するパターンを検出して解析し、意味のある区切りであるシーンを検出するＡＶ信号処理装置のＡＶ信号処理方法において、

前記ＡＶ信号を構成する一連のフレームによって形成されるセグメントの特徴量を抽出する特徴量抽出ステップと、

基準となるセグメントと他のセグメントとの前記特徴量の類似性を測定するための測定基準を算出する算出ステップと、

前記測定基準を用いて、前記基準となるセグメントと前記他のセグメントとの前記類似性を測定する類似性測定ステップと、

類似性測定ステップの処理で測定された前記類似性を用いて、前記基準となるセグメントが前記シーンの境界である可能性を示す測定値を計算する測定値計算ステップと、

前記測定値計算ステップの処理で計算された前記測定値の時間的パターンの変化を解析し、解析結果に基づいて前記基準となるセグメントが前記シーンの境界であるか否かを判定する境界判定ステップと

を含むことを特徴とするＡＶ信号処理方法。

【請求項１２】　供給されたＡＶ信号の内容の意味構造を反映するパターンを検出して解析し、意味のある区切りであるシーンを検出するＡＶ信号処理用のプログラムであって、

前記ＡＶ信号を構成する一連のフレームによって形成されるセグメントの特徴量を抽出する特徴量抽出ステップと、

基準となるセグメントと他のセグメントとの前記特徴量の類似性を測定するための測定基準を算出する算出ステップと、

前記測定基準を用いて、前記基準となるセグメントと前記他のセグメントとの前記類似性を測定する類似性測定ステップと、

類似性測定ステップの処理で測定された前記類似性を用いて、前記基準となるセグメントが前記シーンの境界である可能性を示す測定値を計算する測定値計算ステップと、

前記測定値計算ステップの処理で計算された前記測定値の時間的パターンの変化を解析し、解析結果に基づいて前記基準となるセグメントが前記シーンの境界であるか否かを判定する境界判定ステップと

を含むことを特徴とするコンピュータが読み取り可能なプログラムが記録されている記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、ＡＶ信号処理装置および方法、並びに記録媒体に関し、特に、一連の映像信号の中から所望する部分を選択して再生させる場合に用いて好適なＡＶ

信号処理装置および方法、並びに記録媒体に関する。

【０００２】

【従来の技術】

例えば、ビデオデータに録画されたテレビ番組のような大量の異なる映像データにより構成される映像アプリケーションの中から、興味のある部分等の所望の部分を探して再生したい場合がある。

【０００３】

このように、所望の映像内容を抽出するための一般的な技術としては、アプリケーションの主要場面を描いた一連の映像を並べて作成されたパネルであるストーリボードがある。このストーリボードは、ビデオデータをいわゆるショットに分解し、各ショットにおいて代表される映像を表示したものである。このような映像抽出技術は、そのほとんどが、例えば“G. Ahanger and T.D.C. Little, A survey of technologies for parsing and indexing digital video, J. of Visual Communication and Image Representation 7:28-4, 1996”に記載されているように、ビデオデータからショットを自動的に検出して抽出するものである。

【０００４】

【発明が解決しようとする課題】

ところで、例えば代表的な３０分のテレビ番組中には、数百ものショットが含まれている。そのため、上述した従来の映像抽出技術において、ユーザは、抽出された膨大な数のショットを並べたストーリボードを調べる必要があり、このようなストーリボードを理解するにはユーザに大きな負担を強いる必要があった。

【０００５】

また、従来の映像抽出技術においては、例えば、話し手の変化に応じて交互に２者を撮影した会話場面におけるショットは、冗長のものが多いという問題があった。このように、ショットは、ビデオ構造を抽出する対象としては階層が低すぎて無駄な情報量が多く、このようなショットを抽出する従来の映像抽出技術は、ユーザにとって利便性のよいものではなかった。

【０００６】

また、他の映像抽出技術としては、例えば“A. Merlino, D. Morey and M. Ma

ybury, Broadcast news navigation using story segmentation, Proc. of ACM Multimedia 97, 1997”や特開平１０－１３６２９７号公報に記載されているように、ニュースやフットボールゲームといった特定の内容ジャンルに関する非常に専門的な知識を用いるものがある。しかしながら、この従来の映像抽出技術は、目的のジャンルに関しては良好な結果を得ることができるが、他のジャンルには全く役に立たず、更にジャンルに限定される結果、容易に一般化することができないという問題があった。

【０００７】

さらに、他の映像抽出技術としては、例えば米国特許５７０８７６７号公報に記載されているように、いわゆるストーリユニットを抽出するものがある。しかしながら、この従来の映像抽出技術は、完全に自動化されたものではなく、どのショットが同じ内容を示すものであるかを決定するために、ユーザの操作が必要であった。また、この従来の映像抽出技術は、処理に要する計算が複雑であるとともに、適用対象として映像情報のみに限定されるといった問題もあった。

【０００８】

さらにまた、他の映像抽出技術としては、例えば特開平９－２１４８７９号公報に記載されているように、ショット検出と無音部分検出とを組み合わせることによりシーンを識別するものがある。しかしながら、この従来の映像抽出技術は、無音部分がショット境界に対応した場合のみに限定されたものであった。

【０００９】

また、他の映像抽出技術としては、例えば“H. Aoki, S. Shimotsuji and O. Hori, A shot classification method to select effective key-frames for video browsing, IPSJ Human Interface SIG Notes, 7:43-50, 1996”や特開平９－９３５８８号公報に記載されているように、ストーリボードにおける表示の冗長度を低減する為に、反復された類似ショットを検出するものがある。しかしながら、この従来の映像抽出技術は、映像情報のみに適用できるものであり、音声情報に適用できるものではなかった。

【００１０】

さらに、これら従来技術ではセットトップボックスやデジタルビデオレコーダ

などの家庭機器に実装するにあたり、複数の問題が生じている。それは、主に従来技術では後処理を行うことが前提とされていたためである。具体的には、次の３つの問題が挙げられる。

【００１１】

１つ目の問題は、セグメント数は、コンテンツの長さに依存し、一定であってもその中に含まれるショットの数が一定でない。そのためシーン検出に必要なメモリ量の固定ができないので必要とするメモリ量を過剰に設定しなければならなかった。これはメモリ量の少ない家庭機器では大きな問題であった。

【００１２】

２つ目の問題は、家庭機器では、決められた時間内に決められた処理を必ず終わらせなければならない実時間処理が必要とされる。しかし、セグメント数が固定できなく、また、後処理処理を行わなければならないため、常に決められた時間内に処理を終わらせるのは困難であった。このことは家庭用機器に実装されている高性能でないCPUを使用しなけらばならない場合、さらに実時間処理を行うことが困難であることを意味する。

【００１３】

３つ目の問題は、今まで述べてきたように後処理処理が必要であるため、セグメントが生成される毎にシーン検出の処理結果が終わらせることができない。これは録画途中で何らかの理由で録画状態が止まった場合、それまでの途中結果を得られないことを意味する。これは録画しながら逐次処理ができないことを意味し、家庭用機器では大きな問題になる。

【００１４】

また、従来技術では、シーンを決定する場合、セグメントの繰り返しのパターンやそれ以外のセグメントのグループ化などによる方法を用いていたためシーンの検出結果は一意的になっていた。故に検出された境界が実際のシーンの境界である可能性が高いか低いかを判断することは不可能であり、段階的にシーンの検出数を制御することができなかった。

【００１５】

さらに、ビデオを一覧するに当たって、見易くするため得られたシーンの数を

出来る限り少なくすることが必要となる。そのゆえに、検出したシーンの数が限定された場合に、どのシーンを見せるとよいかという問題が生じる。そのため、得られたシーンの各々の重要性が解れば、その重要性の順番に従い、シーンを見せると一覧するためによい。ただし、従来技術では得られたシーンがどの程度重要であるかを計る尺度を提供していない。

【００１６】

本発明はこのような状況に鑑みてなされたものであり、録画したビデオデータを任意のシーンから再生できるように、シーンの境界を検出することを目的とする。

【００１７】

【課題を解決するための手段】

本発明のＡＶ信号処理装置は、ＡＶ信号を構成する一連のフレームによって形成されるセグメントの特徴量を抽出する特徴量抽出手段と、基準となるセグメントと他のセグメントとの特徴量の類似性を測定するための測定基準を算出する算出手段と、測定基準を用いて、基準となるセグメントと他のセグメントとの類似性を測定する類似性測定手段と、類似性測定手段が測定した類似性を用いて、基準となるセグメントがシーンの境界である可能性を示す測定値を計算する測定値計算手段と、測定値計算手段が計算した測定値の時間的パターンの変化を解析し、解析結果に基づいて基準となるセグメントがシーンの境界であるか否かを判定する境界判定手段とを含むことを特徴とする。

【００１８】

ＡＶ信号には、映像信号および音声信号のうちの少なくとも一方を含むようにすることができる。

【００１９】

本発明のＡＶ信号処理装置は、基準となるセグメントに対応する測定値の変化の程度を示す強度値を計算する強度値計算手段をさらに含むことができる。

【００２０】

前記測定値計算手段には、基準となるセグメントに対して、所定の時間領域内における類似セグメントを求め、類似セグメントの時間分布を解析し、過去と未

来に存在する比率を数値化して測定値を計算させるようにすることができる。

【００２１】

前記境界判定手段には、測定値の絶対値の総和にも基づき、基準となるセグメントがシーンの境界であるか否かを判定させるようにすることができる。

【００２２】

本発明のＡＶ信号処理装置は、ＡＶ信号に映像信号が含まれる場合、映像セグメントの基本単位となるショットを検出して、音声セグメントを生成する音声セグメント生成手段をさらに含むことができる。

【００２３】

本発明のＡＶ信号処理装置は、ＡＶ信号に音声信号が含まれる場合、音声信号の特徴量および無音区間のうちの少なくとも一方を用いて、音声セグメントを生成する音声セグメント生成手段をさらに含むことができる。

【００２４】

映像信号の特徴量には、少なくともカラーヒストグラムが含まれるようにすることができる。

【００２５】

音声信号の特徴量には、音量およびスペクトラムのうちの少なくとも一方が含まれるようにすることができる。

【００２６】

前記境界判定手段には、予め設定され閾値と測定値を比較することにより、基準となるセグメントがシーンの境界であるか否かを判定させるようにすることができる。

【００２７】

本発明のＡＶ信号処理方法は、ＡＶ信号を構成する一連のフレームによって形成されるセグメントの特徴量を抽出する特徴量抽出ステップと、基準となるセグメントと他のセグメントとの特徴量の類似性を測定するための測定基準を算出する算出ステップと、測定基準を用いて、基準となるセグメントと他のセグメントとの類似性を測定する類似性測定ステップと、類似性測定ステップの処理で測定された類似性を用いて、基準となるセグメントがシーンの境界である可能性を示

す測定値を計算する測定値計算ステップと、測定値計算ステップの処理で計算された測定値の時間的パターンの変化を解析し、解析結果に基づいて基準となるセグメントがシーンの境界であるか否かを判定する境界判定ステップとを含むことを特徴とする。

【００２８】

本発明の記録媒体のプログラムは、ＡＶ信号を構成する一連のフレームによって形成されるセグメントの特徴量を抽出する特徴量抽出ステップと、基準となるセグメントと他のセグメントとの特徴量の類似性を測定するための測定基準を算出する算出ステップと、測定基準を用いて、基準となるセグメントと他のセグメントとの類似性を測定する類似性測定ステップと、類似性測定ステップの処理で測定された類似性を用いて、基準となるセグメントがシーンの境界である可能性を示す測定値を計算する測定値計算ステップと、測定値計算ステップの処理で計算された測定値の時間的パターンの変化を解析し、解析結果に基づいて基準となるセグメントがシーンの境界であるか否かを判定する境界判定ステップとを含むことを特徴とする。

【００２９】

本発明のＡＶ信号処理装置および方法、並びに記録媒体のプログラムにおいては、ＡＶ信号を構成する一連のフレームによって形成されるセグメントの特徴量が抽出され、基準となるセグメントと他のセグメントとの特徴量の類似性を測定するための測定基準が算出され、測定基準を用いて、基準となるセグメントと他のセグメントとの類似性が測定され、測定された類似性を用いて、基準となるセグメントがシーンの境界である可能性を示す測定値が計算される。また、計算された測定値の時間的パターンの変化が解析され、解析結果に基づいて基準となるセグメントがシーンの境界であるか否かが判定される。

【００３０】

【発明の実施の形態】

本発明は、ビデオデータを意味的なセグメントのまとまりであるシーンという単位に切り分けることが目的である。この切り分けるという意味はシーンとシーンの境界を検出するということである。シーンを構成するセグメントはシーンそ

れぞれに固有な特徴を持っているため、シーンが隣接する場合、その境界を超えると構成しているセグメントの特徴に顕著な違いが現れる。換言すれば、そのような顕著な違いが現れるところがシーンの境界であり、それを検出することにより一連のセグメントをシーン単位に切り分けることが可能になる。

【００３１】

この処理を行うに当たり、上述した従来技術と同ように、最初に対象となるビデオデータをセグメント単位に分割する。分割して得たセグメントは時系列を成し、各セグメントとその次のセグメントとの間にシーン境界があるのかを判断することが必要となる。各セグメントを基準とし、その近隣のセグメントの中に似ているセグメントが時間的に何処にあるのかを調べる。

【００３２】

もしシーン境界があった場合、過去に集中して存在していたパターンから、未来に集中して存在するパターンへと短い時間で特異な変化が現れる変化点が検出される。その変化点から次の変化点までが一つのシーンである。このようなパターンの変化が起こるところを見つけるため、シーンの境界の前後で局所的な変化を見るだけで十分な情報が得られる。

【００３３】

さらにこの局所的変化の大きさの大小を測定することによりシーンの切り分けを段階的に制御することも可能である。これは視覚的な変化点がシーンの意味的な変化点と良く一致することが経験的に判明したことからである。本発明は以上のことを基本にしてシーンの境界を検出し、ビデオデータなどのシーンを切り分けるためのものである。またこのシーン境界情報をもとにビデオデータを見やすく表示することを可能とする。

【００３４】

次に、本発明の概要を具体的に説明する。まず、シーンとシーンの境界が存在する場合と存在しない場合に分けて、それぞれの特徴について説明する。あるビデオデータの具体例を図２に示す。同図では、ビデオデータの単位はセグメント単位で示されており、３つのシーン１乃至シーン３によって構成されているものである。同図において時間軸は右方向に向いているものとする。境界が存在しな

い領域を非境界領域とし、境界が存在している領域を境界領域とし、図４に詳細に示してある。

【００３５】

図４（Ａ）の非境界領域に示してあるのはシーン２の時間内の部分であり、他のシーンとの境界が存在していないセグメント３乃至セグメント１１の時間領域である。また、これと対照的に図４（Ｂ）の境界領域はシーン２とシーン３の境界領域を含むところでシーンとシーンの隣接しているセグメント８乃至セグメント１５の時間領域を示している。

【００３６】

まず、境界が存在しない場合を表している非境界領域の特徴について説明する。非境界領域は、類似したセグメントだけで構成されているので、非境界領域の中の基準セグメントに対して過去、未来の時間帯と分けた場合ほぼ均等に類似セグメントは存在する。そのため類似セグメントの分布パターンには特異な変化のパターンは現れない。

【００３７】

境界領域は、非境界領域と異なり、２つのシーンが隣接している境界点を含む時間帯の部分を表している。ここでシーンというのは互いに高い類似性を持ったセグメントからなっているものを意味する。そのため、シーン２を構成しているセグメント８乃至セグメント１１と、異なるシーン３を構成しているセグメント１２乃至セグメント１５とが隣り合っており、それらの境界を挟んでシーンのセグメントの特徴がそれぞれ変る。

【００３８】

シーンの境界を検出するには、まず各セグメントを時間的基準(現在)と仮定する。それぞれに対し、最も類似したセグメントの時間的分布パターン(基準から見て過去であるのか未来であるのか)の変化を調べることにより実現できる。

【００３９】

これは図４（Ｂ）に示す境界領域からわかるように、セグメント８乃至セグメント１１が順に時間的基準となって境界に近づくにつれ、最も類似なセグメントが未来に対して過去に存在する比率が高くなって行き、境界直近(シーンの終り)

では１００％になる。そして境界を越えた直後(次のシーンの先頭)では過去に対して未来に存在する比率が１００％になり、セグメント１２乃至セグメント１５が順に時間的基準となるにつれ、その比率が低くなって行く。

【００４０】

したがって、このような最も類似なセグメントの時間分布比率のパターンの変化より、シーンの境界の可能性の高い場所を特定できる。また、この典型的なパターンはシーンの境界付近の局所的な部分に現れる確率が非常に高いので、境界近辺だけを調べればそのパターンの変化から境界を特定できる。これは言い換えれば、類似セグメントの分布パターンを調べる時間領域を必要以上に大きく取らなくても良いということになる。

【００４１】

また、これらのパターンの変化を数値化すると、その値の変化の度合いがシーンの視覚的変化の度合いに連動している。そしてシーンの視覚的変化の度合いはシーンの意味的な変化の度合いに連動していることが経験上および実験的結果によってわかっている。したがってこの数値化した値を境界性測定値とすると、この値の大小によりシーンの意味的度合いの大小に対応したシーンを検出することが可能となる。

【００４２】

次に、本発明の一実施の形態である映像音声処理装置について説明するが、その前に、映像音声処理装置が処理の対象とするビデオデータについて説明する。

【００４３】

本発明においては、処理対象とするビデオデータを、図１に示すようにモデル化し、フレーム、セグメント、シーンの３つのレベルに階層化されたデータ構造を有するものとする。すなわち、ビデオデータは、最下位層において、一連のフレームにより構成される。また、ビデオデータは、フレームの１つ上の階層として、連続するフレームのひと続きから形成されるセグメントにより構成される。さらに、ビデオデータは、最上位層において、このセグメントを意味のある関連に基づきまとめて形成されるシーンにより構成される。

【００４４】

このビデオデータは、一般に、映像および音声の両方の情報を含む。すなわち、このビデオデータにおいてフレームは、単一の静止画像である映像フレームと、一般に数十乃至数百ミリセカンド／長といった短時間においてサンプルされた音声情報を表す音声フレームが含まれる。

【００４５】

また、映像セグメントは、単一のカメラにより連続的に撮影された一連の映像フレームから構成され、一般にはショットと呼ばれる。

【００４６】

一方、音声セグメントについては、多くの定義が可能であり、例として次に示すようなものが考えられる。音声セグメントは、一般によく知られている方法により検出されたビデオデータ中の無音期間により境界を定められて形成されるものがある。また、音声セグメントは、“D. Kimber and L. Wilcox, Acoustic Segmentation for Audio Browsers, Xerox Parc Technical Report”に記載されているように、例えば、音声、音楽、ノイズ、無音等のように少数のカテゴリに分類された音声フレームのひと続きから形成されるものがある。さらに、音声セグメントは、“S. Pfeiffer, S. Fischer and E. Wolfgang, Automatic Audio Content Analysis, Proceeding of ACM Multimedia 96, Nov. 1996, pp21-30”に記載されているように、２枚の連続する音声フレーム間のある特徴における大きな変化を音声の変わり目として検出し、これに基づいて決定される場合もある。

【００４７】

シーンは、ビデオデータの内容を意味に基づくより高いレベルのものである。シーンは、主観的なものであり、ビデオデータの内容あるいはジャンルに依存する。シーンは、その特徴が互いに類似性を示す映像セグメントまたは音声セグメントで構成されている。

【００４８】

ここでは、ビデオデータ内の各セグメントについて、その近隣に存在する類似的特徴を持っているセグメントが、過去に集中して存在していたパターンから、未来に集中して存在するパターンへと特異な変化を示す変化点を検出し、その変化点から次の変化点を一つのシーンとするものである。このようなパターンがシ

ーンの切れ目と対応するのは、各シーンに含まれているセグメントの特徴が異なるためにシーンの境界でセグメントの類似的特徴が大きく変化するからである。これはビデオデータにおける高いレベルでの意味のある構造と非常に関係があり、シーンは、このようなビデオデータにおける高いレベルでの意味を持ったまとまりを示すものである。

【００４９】

次に、本発明の一実施の形態である映像音声処理装置の構成例について、図３を参照して説明する。映像音声処理装置は、上述したビデオデータにおけるセグメントの特徴量を用いてセグメント間の類似性を測定し、これらのセグメントをシーンにまとめてビデオ構造を自動的に抽出するものであり、映像セグメントおよび音声セグメントの両方に適用できるものである。

【００５０】

映像音声処理装置は、図３に示すように、入力されるビデオデータのストリームを映像または音声、あるいは両方のセグメントに分割するビデオ分割部１１、ビデオデータの分割情報を記憶するビデオセグメントメモリ１２、各映像セグメントにおける特徴量を抽出する映像特徴量抽出部１３、各音声セグメントにおける特徴量を抽出する音声特徴量抽出部１４、映像セグメントおよび音声セグメントの特徴量を記憶するセグメント特徴量メモリ１５、映像セグメントおよび音声セグメントをシーンにまとめるシーン検出部１６、および、２つのセグメント間の類似性を測定する特徴量類似性測定部１７より構成される。

【００５１】

ビデオ分割部１１は、入力される、例えば、MPEG(Moving Picture Experts Group)１、MPEG２、またはいわゆるＤＶ(Digital Video)などの圧縮ビデオデータフォーマットを含む種々のディジタル化されたフォーマットにおける映像データと音声データとからなるビデオデータのストリームを映像、音声またはこれらの両方のセグメントに分割するものである。

【００５２】

ビデオ分割部１１は、入力されるビデオデータが圧縮フォーマットであった場合、この圧縮ビデオデータを完全伸張することなく直接処理することができる。

ビデオ分割部１１は、入力されたビデオデータを処理し、映像セグメントと音声セグメントとに分割する。また、ビデオ分割部１１は、入力したビデオデータを分割した結果である分割情報を後段のビデオセグメントメモリ１２に出力する。さらに、ビデオ分割部１１は、映像セグメントと音声セグメントとに応じて、分割情報を後段の映像特徴量抽出部１３および音声特徴量抽出部１４に出力する。

【００５３】

ビデオセグメントメモリ１２は、ビデオ分割部１１から供給されたビデオデータの分割情報を記憶する。また、ビデオセグメントメモリ１２は、後述するシーン検出部１６からの問い合わせに応じて、分割情報をシーン検出部１６に出力する。

【００５４】

映像特徴量抽出部１３は、ビデオ分割部１１によりビデオデータを分割して得た各映像セグメント毎の特徴量を抽出する。映像特徴量抽出部１３は、圧縮映像データを完全伸張することなく直接処理することができる。映像特徴量抽出部１３は、抽出した各映像セグメントの特徴量を後段のセグメント特徴量メモリ１５に出力する。

【００５５】

音声特徴量抽出部１４は、ビデオ分割部１１によりビデオデータを分割して得た各音声セグメント毎の特徴量を抽出する。音声特徴量抽出部１４は、圧縮音声データを完全伸張することなく直接処理することができる。音声特徴量抽出部１４は、抽出した各音声セグメントの特徴量を後段のセグメント特徴量メモリ１５に出力する。

【００５６】

セグメント特徴量メモリ１５は、映像特徴量抽出部１３および音声特徴量抽出部１４からそれぞれ供給された映像セグメントおよび音声セグメントの特徴量を記憶する。セグメント特徴量メモリ１５は、後述する特徴量類似性測定部１７からの問い合わせに応じて、記憶している特徴量やセグメントを特徴量類似性測定部１７に出力する。

【００５７】

シーン検出部１６は、ビデオセグメントメモリ１２に保持された分割情報と、セグメント間の類似性とを用いて、映像セグメントおよび音声セグメントがシーンの境界であるかを判断する。シーン検出部１６は、各セグメントの近隣の最も類似な特徴量を持つセグメントの分布パターンが、過去に集中した状態から未来に集中した状態へ切り替わる変化点を特定することにより、シーンの境界を検出し先頭部と最後部を確定する。シーン検出部１６は、セグメントが発生する毎に1セグメント分、時系列的に移動させ、近隣の最も類似しているセグメントの分布パターンを測定する。シーン検出部１６は、特徴量類似性測定部１７を用いて、近隣のセグメントで最も類似しているものの数を特定する。すなわち、特徴空間における特徴量の最近傍の数を求める。そしてセグメントの最近傍の類似セグメントがそのセグメントを境にして過去に存在するものと未来に存在するものの個数の違いのパターンの変化からシーンの境界を特定する。

【００５８】

特徴量類似性測定部１７は、各セグメントとその近隣のセグメントとの類似性を測定する。特徴量類似性測定部１７は、あるセグメントに関する特徴量を検索するようにセグメント特徴量メモリ１５に問いかける。

【００５９】

ビデオデータ記録部１８は、ビデオストリームおよびビデオデータに関する各種のデータである、いわゆる付加情報データを記録する。ここにシーン検出部１６から出力されたシーン境界情報およびシーンに対して計算された強度値が保存される。

【００６０】

ビデオ表示部１９は、ビデオデータ記録部１８からのビデオデータを、各種付加情報データに基き、サムネイルのような表示方法やランダムアクセス方法などを実現する。これはユーザの視聴方法に自由度を増やし、利便性良くビデオデータを表示する。

【００６１】

制御部２０は、ドライブ２１を制御して、磁気ディスク２２、光ディスク２３、光磁気ディスク２４、または半導体メモリ２５に記憶されている制御用プログ

ラムを読み出し、読み出した制御用プログラムに基づいて、映像音声処理装置の各部を制御する。

【００６２】

映像音声処理装置は、図５に概略を示すような一連の処理を行うことによって、シーンを検出する。

【００６３】

まず、映像音声処理装置は、同図に示すように、ステップＳ１において、ビデオ分割を行う。すなわち映像音声処理装置は、ビデオ分割部１１に入力されたビデオデータを映像セグメントまたは音声セグメントのいずれか、あるいは可能であればその両方に分割する。

【００６４】

映像音声処理装置が適用するビデオ分割方法には、特に前提要件を設けない。例えば、映像音声処理装置は、“G. Ahanger and T.D.C. Little, A survey of technologies for parsing and indexing digital video, J. of Visual Communication and Image Representation 7:28-4, 1996”に記載されているような方法によりビデオ分割を行う。このようなビデオ分割の方法は、当該技術分野ではよく知られたものであり、映像音声処理装置は、いかなるビデオ分割方法も適用できるものとする。

【００６５】

次に、映像音声処理装置は、ステップＳ２において、特徴量の抽出を行う。すなわち映像音声処理装置は、映像特徴量抽出部１３や音声特徴量抽出部１４により、そのセグメントの特徴を表す特徴量を計算する。映像音声処理装置では、例えば、各セグメントの時間長や、カラーヒストグラムやテクスチャフィーチャといった映像特徴量や、周波数解析結果、レベル、ピッチといった音声特徴量やアクティビティ測定結果等が、適用可能な特徴量として計算される。勿論、映像音声処理装置は、適用可能な特徴量としてこれらに限定されるものではない。

【００６６】

続いて、映像音声処理装置は、ステップＳ３において、特徴量を用いたセグメントの類似性測定を行う。すなわち映像音声処理装置は、特徴量類似性測定部１

７により非類似性測定を行い、その測定基準により、セグメントとその近隣のセグメントがどの程度類似しているかを測定する。映像音声処理装置は、先のステップＳ２において抽出した特徴量を用いて、非類似性測定基準を計算する。

【００６７】

そして、映像音声処理装置は、ステップＳ４において、セグメントがシーンの切れ目にあたるか否かを判断する。すなわち、映像音声処理装置は、先のステップＳ３において計算した非類似性測定基準と、先のステップＳ２において抽出した特徴量とを用いて、各セグメントを現在と見なし、近接の類似したセグメントが、その基準とするセグメントに対し過去か未来かどちらに存在比率が高いかを求め、その存在比の率変化のパターンを調べ、シーンの境界であるか否かの判断をする。映像音声処理装置は、このようにして最終的に各セグメントがシーンの切れ目であるか否かを出力する。

【００６８】

このような一連の処理を経ることによって、映像音声処理装置は、ビデオデータからシーンを検出することができる。

【００６９】

したがって、ユーザは、この結果を用いることによって、ビデオデータの内容を要約したり、ビデオデータ中の興味のあるポイントに迅速にアクセスしたりすることが可能となる。

【００７０】

以下、上述した一連の処理を各ステップの処理毎に、より詳細に説明する。

【００７１】

ステップＳ１におけるビデオ分割について説明する。映像音声処理装置は、ビデオ分割部１１に入力されたビデオデータを映像セグメントまたは音声セグメントのいずれか、あるいは可能であればその両方に分割するが、このビデオデータにおけるセグメントの境界を自動的に検出するための技術は多くのものがあり、映像音声処理装置において、このビデオ分割方法に特別な前提要件を設けないことは上述した通りである。

【００７２】

一方、映像音声処理装置において、後の処理によるシーン検出の精度は、本質的に、基礎となるビデオ分割の精度に依存する。なお、映像音声処理装置におけるシーン検出は、ある程度ビデオ分割時のエラーを許容することができる。特に、映像音声処理装置において、ビデオ分割は、セグメント検出が不十分である場合よりも、セグメント検出を過度に行う場合の方が好ましい。映像音声処理装置は、類似したセグメントの検出が過度である結果である限り、一般に、シーン検出の際に検出過度であるセグメントを同一シーンとしてまとめることができる。

【００７３】

ステップＳ２における特徴量抽出について説明する。特徴量とは、セグメントの特徴を表すとともに、異なるセグメント間の類似性を測定するためのデータを供給するセグメントの属性である。映像音声処理装置は、映像特徴量抽出部１３や音声特徴量抽出部１４において各セグメントの特徴量を計算し、セグメントの特徴を表す。

【００７４】

映像音声処理装置は、いかなる特徴量の具体的詳細にも依存するものではないが、映像音声処理装置において用いて効果的であると考えられる特徴量としては、例えば以下に示す映像特徴量、音声特徴量、映像音声共通特徴量のようなものがある。映像音声処理装置において適用可能となるこれら特徴量の必要条件は、非類似性の測定が可能であることである。また映像音声処理装置は、効率化のために、特徴量抽出と上述したビデオ分割とを同時に行うことがある。以下に説明する特徴量は、このような処理を可能にするものである。

【００７５】

上記特徴量としては、まず映像に関するものが挙げられる。以下では、これを映像特徴量と称することにする。映像セグメントは、連続する映像フレームにより構成されるため、映像セグメントから適切な映像フレームを抽出することによって、その映像セグメントの描写内容を、抽出した映像フレームで特徴付けることが可能である。すなわち映像セグメントの類似性は、適切に抽出された映像フレームの類似性で代替可能である。つまり映像特徴量は、映像音声処理装置で用いることができる重要な特徴量の１つである。この場合の映像特徴量は、単独で

は静的な情報しか表せないが、映像音声処理装置は、後述するような方法を適用することによって、この映像特徴量に基づく映像セグメントの動的な特徴を抽出する。

【００７６】

映像特徴量として既知のものは多数存在するが、シーン検出のためには以下に示す色特徴量（ヒストグラム）および映像相関が、計算コストと精度との良好な兼ね合いを与えることを見出したことから、映像音声処理装置は、映像特徴として、色特徴量および映像相関を用いることにする。

【００７７】

映像音声処理装置において、映像における色は、２つの映像が類似しているかを判断する際の重要な材料となる。カラーヒストグラムを用いて映像の類似性を判断することは、例えば“G. Ahanger and T.D.C. Little, A survey of technologies for parsing and indexing digital video, J. of Visual Communication and Image Representation 7:28-4, 1996”に記載されているように、よく知られている。

【００７８】

ここでカラーヒストグラムとは、例えばLUVやRGB等の３次元色空間をｎ個の領域に分割し、映像における画素の、各領域での出現頻度の相対的割合を計算したものである。そして、得られた情報からは、ｎ次元ベクトルが与えられる。圧縮されたビデオデータについては、例えば米国特許５７０８７６７号公報に記載されているように、カラーヒストグラムを、圧縮データから直接抽出することができる。

【００７９】

映像音声処理装置では、セグメントを構成する映像（MPEG1／2, DVなど一般的に使われている方式）における元々のYUV色空間のヒストグラムベクトルを得る。

【００８０】

映像音声処理装置では、セグメントを構成する映像（MPEG1／2, DVなど一般的に使われている方式）における元来のYUV色空間を、色チャンネル当たり２ビッ

トでサンプリングして構成した、$2^{2\cdot3}$＝６４次元のヒストグラムベクトルを得る。

【００８１】

このようなヒストグラムは、映像の全体的な色調を表すが、これには時間情報が含まれていない。そこで、映像音声処理装置では、もう一つの映像特徴量として、映像相関を計算する。映像音声処理装置でのシーン検出において、複数の類似セグメントが互いに交差した構造は、それがまとまった１つのシーン構造であることの有力な指標となる。

【００８２】

例えば会話場面において、カメラの位置は、２人の話し手の間を交互に移動するが、カメラは通常、同一の話し手を再度撮影するときには、ほぼ同じ位置に戻る。このような場合における構造を検出するためには、グレイスケールの縮小映像に基づく相関がセグメントの類似性の良好な指標となることを見出したことから、映像音声処理装置では、元の映像をＭ×Ｎの大きさのグレイスケール映像に間引き縮小し、これを用いて映像相関を計算する。ここで、ＭとＮは、両方とも小さな値で十分であり、例えば８×８である。つまり、これらの縮小グレイスケール映像は、ＭＮ次元の特徴量ベクトルとして解釈される。

【００８３】

さらに上述した映像特徴量とは異なる特徴量としては、音声に関するものが挙げられる。以下では、この特徴量を音声特徴量と称することにする。音声特徴量とは、音声セグメントの内容を表すことができる特徴量であり、映像音声処理装置は、この音声特徴量として、周波数解析、ピッチ、レベル等をを用いることができる。これらの音声特徴量は、種々の文献により知られているものである。

【００８４】

まず、映像音声処理装置は、フーリエ変換等の周波数解析を行うことによって、単一の音声フレームにおける周波数情報の分布を決定することができる。映像音声処理装置は、例えば、１つの音声セグメントにわたる周波数情報の分布を表すために、FFT（Fast Fourier Transform；高速フーリエ変換）成分、周波数ヒストグラム、パワースペクトル、ケプストラム(Cepstrum)、その他の特徴量を用

いることができる。

【００８５】

また、映像音声処理装置は、平均ピッチや最大ピッチなどのピッチや、平均ラウドネスや最大ラウドネスなどの音声レベルもまた、音声セグメントを表す有効な音声特徴量として用いることができる。

【００８６】

さらに他の特徴量としては、映像音声共通特徴量が挙げられる。これは、特に映像特徴量でもなく音声特徴量でもないが、映像音声処理装置において、シーン内のセグメントの特徴を表すのに有用な情報を与えるものである。映像音声処理装置は、この映像音声共通特徴量として、セグメント長とアクティビティとを用いる。

【００８７】

映像音声処理装置は、映像音声共通特徴量として、セグメント長を用いることができる。このセグメント長は、セグメントにおける時間長である。一般に、シーンは、そのシーンに固有のリズム特徴を有する。このリズム特徴は、シーン内のセグメント長の変化として表れる。例えば、迅速に連なった短いセグメントは、コマーシャルを表す。一方、会話シーンにおけるセグメントは、コマーシャルの場合よりも長く、また会話シーンには、相互に組み合わされたセグメントが互いに類似しているという特徴がある。映像音声処理装置は、このような特徴を有するセグメント長を映像音声共通特徴量として用いることができる。

【００８８】

また、映像音声処理装置は、映像音声共通特徴量として、アクティビティを用いることができる。アクティビティとは、セグメントの内容がどの程度動的あるいは静的であるように感じられるかを表す指標である。例えば、視覚的に動的である場合、アクティビティは、カメラが対象物に沿って迅速に移動する度合若しくは撮影されているオブジェクトが迅速に変化する度合を表す。

【００８９】

このアクティビティは、カラーヒストグラムのような特徴量のフレーム間非類似性の平均値を測定することにより、間接的に計算される。ここで、フレームｉ

とフレームｊとの間で測定された特徴量Ｆに対する非類似性測定基準をｄＦ（ｉ，ｊ）と定義すると、映像アクティビティＶＦは、次式（１）のように定義される。

【数１】

$$V_F = \frac{\sum_{i=b}^{f-1} d_F(i, i+1)}{f-b+1}$$

【００９０】

式（１）において、ｂとｆはそれぞれ、１セグメントにおける最初と最後のフレームのフレーム番号である。映像音声処理装置は、具体的には、例えば上述したヒストグラムを用いて、映像アクティビティＶＦを計算する。

【００９１】

ところで、上述した映像特徴量を始めとする特徴量は、基本的にはセグメントの静的情報を表すものであることは上述した通りであるが、セグメントの特徴を正確に表すためには、その動的情報も考慮する必要がある。そこで、映像音声処理装置は、以下に示すような特徴量のサンプリング方法により動的情報を表す。

【００９２】

映像音声処理装置は、例えば図５に示すように、１セグメント内の異なる時点から１以上の静的な特徴量を抽出する。このとき、映像音声処理装置は、特徴量の抽出数を、そのセグメント表現における忠実度の最大化とデータ冗長度の最小化とのバランスをとることにより決定する。例えば、セグメント内のある１画像が当該セグメントのキーフレームとして指定可能な場合には、そのキーフレームから計算されたヒストグラムが、抽出すべきサンプリング特徴量となる。

【００９３】

映像音声処理装置は、後述するサンプリング方法を用いて、対象とするセグメントにおいて、特徴として抽出可能なサンプルのうち、どのサンプルを選択するかを決定する。

【００９４】

ところで、あるサンプルが常に所定の時点、例えばセグメント内の最後の時点

において選択される場合を考える。この場合、黒フレームへ変化してゆく（フェードしてゆく）任意の２つのセグメントについては、サンプルが同一の黒フレームとなるため、同一の特徴量が得られる結果になる恐れがある。すなわち、これらのセグメントの映像内容がいかなるものであれ、選択した２つのフレームは、極めて類似していると判断されてしまう。このような問題は、サンプルが良好な代表値でないために発生するものである。

【００９５】

そこで、映像音声処理装置は、このように固定点で特徴量を抽出するのではなく、セグメント全体における統計的な代表値を抽出することとする。ここでは、一般的な特徴量のサンプリング方法を２つの場合、すなわち、特徴量を実数のｎ次元ベクトルとして表すことができる第１の場合と、非類似性測定基準しか利用できない第２の場合とについて説明する。なお、第１の場合は、ヒストグラムやパワースペクトル等、最もよく知られている映像特徴量および音声特徴量が含まれる。

【００９６】

第１の場合においては、サンプル数は、事前にｋと決められており、映像音声処理装置は、“L. Kaufman and P.J. Rousseeuw, Finding Groups in Data:An Introduction to Cluster Analysis, John-Wiley and sons, 1990”に記載されてよく知られているｋ平均値クラスタリング法(k-means-clustering method)を用いて、セグメント全体についての特徴量をｋ個の異なるグループに自動的に分割する。そして、映像音声処理装置は、サンプル値として、ｋ個の各グループから、グループの重心値（centroid）またはこの重心値に近いサンプルを選択する。映像音声処理装置におけるこの処理の複雑度は、サンプル数に関して単に直線的に増加するに留まる。

【００９７】

一方、第２の場合においては、映像音声処理装置は、“L. Kaufman and P.J. Rousseeuw, Finding Groups in Data:An Introduction to Cluster Analysis, John-Wiley and sons, 1990”に記載されているｋ－メドイドアルゴリズム法(k-medoids algorithm method)を用いて、ｋ個のグループを形成する。そして、映像

音声処理装置は、サンプル値として、ｋ個の各グループ毎に、上述したグループのメドイド(medoid)を用いる。

【００９８】

なお、映像音声処理装置においては、抽出された動的特徴を表す特徴量についての非類似性測定基準を構成する方法は、その基礎となる静的な特徴量の非類似性測定基準に基づくが、これについては後述する。

【００９９】

このようにして、映像音声処理装置は、静的な特徴量を複数抽出し、これら複数の静的な特徴量を用いることで、動的特徴を表すことができる。

【０１００】

以上のように、映像音声処理装置は、種々の特徴量を抽出することができる。これらの各特徴量は、一般に、単一ではセグメントの特徴を表すのに不十分であることが多い。そこで、映像音声処理装置は、これらの各種特徴量を組み合わせることで、互いに補完し合う特徴量の組を選択することができる。例えば、映像音声処理装置は、上述したカラーヒストグラムと映像相関とを組み合わせることによって、各特徴量が有する情報よりも多くの情報を得ることができる。

【０１０１】

次に、図５のステップＳ３における特徴量を用いたセグメントの類似性測定について説明する。映像音声処理装置は、２つの特徴量について、それがどの程度非類似であるかを測定する実数値を計算する関数である非類似性測定基準を用いて、特徴量類似性測定部１７によりセグメントの類似性測定を行う。この非類似性測定基準は、その値が小さい場合は２つの特徴量が類似していることを示し、値が大きい場合は非類似であることを示す。ここでは、特徴量Ｆに関する２つのセグメントS_1，S_2の非類似性を計算する関数を非類似性測定基準ｄＦ（S_1，S_2）と定義する。なお、この関数は、以下の式（２）で与えられる関係を満足させる必要がある。

【数２】

$d_F(S_1, S_2) = 0$ ($S_1 = S_2$のとき)
$d_F(S_1, S_2) \geqq 0$ (全てのS_1, S_2について)
$d_F(S_1, S_2) = d_F(S_2, S_1)$ (全てのS_1, S_2について)

【０１０２】

ところで、非類似性測定基準の中には、ある特定の特徴量にのみ適用可能なものがあるが、“G. Ahanger and T.D.C. Little, A survey of technologies for parsing and indexing digital video, J. of Visual Communication and Image Representation 7:28-4, 1996”や“L. Kaufman and P.J. Rousseeuw, Finding Groups in Data:An Introduction to Cluster Analysis, John-Wiley and sons, 1990”に記載されているように、一般には、多くの非類似性測定基準は、ｎ次元空間における点として表される特徴量についての類似性を測定するのに適用可能である。

【０１０３】

その具体例は、ユークリッド距離、内積、Ｌ１距離等である。ここで、特にＬ１距離が、ヒストグラムや映像相関などの特徴量を含む種々の特徴量に対して有効に作用することから、映像音声処理装置は、Ｌ１距離を導入する。ここで、２つのｎ次元ベクトルをＡ，Ｂとした場合、Ａ，Ｂ間のＬ１距離ｄＬ１（Ａ，Ｂ）は次式（３）で与えられる。

【数３】

$$d_{L1} = (A, B) = \sum_{i=1}^{n} |Ai - Bi$$

ここで下付文字ｉは、ｎ次元ベクトルＡ，Ｂそれぞれのｉ次元の要素を示すものである。

【０１０４】

また、映像音声処理装置は、上述したように、動的特徴を表す特徴量として、セグメントにおけるよう々な時点での静的な特徴量を抽出する。そして、映像音声処理装置は、抽出された二つの動的特徴量間の類似性を決定するために、その非類似性測定基準として、その基礎となる静的特徴量の間の非類似性測定基準を

用いる。これら動的特徴量の非類似性測定基準は、多くの場合、各動的特徴量から選択された最も類似した静的特徴量の対の非類似性値を用いて決定されるのが最良である。この場合、２つの抽出された動的特徴量ＳＦ$_1$，ＳＦ$_2$の間の非類似性測定基準は、次式（４）のように定義される。

【数４】

$$d_S(SF_1, SF_2) = \min_{F_1 \in SF_1, F_2 \in SF_2} d_F(F_1, F_2)$$

【０１０５】

ここで、上式（４）における関数ｄＦ（Ｆ$_1$，Ｆ$_2$）は、その基礎となる静的特徴量Ｆについての非類似性測定基準を示す。なお、場合によっては、特徴量の非類似性の値の最小をとる代わりに、最大値または平均値をとってもよい。

【０１０６】

ところで、映像音声処理装置は、セグメントの類似性を決定する上で、単一の特徴量だけでは不十分であり、同一セグメントに関する多数の特徴量からの情報を組み合わせることを必要とする場合も多い。この１つの方法として、映像音声処理装置は、種々の特徴量に基づく非類似性を、それぞれの特徴量の重み付き組み合わせとして計算する。すなわち、映像音声処理装置は、ｋ個の特徴量Ｆ$_1$，Ｆ$_2$，・・・，Ｆ$_k$が存在する場合、次式（５）に表される組み合わせた特徴量に関する非類似性測定基準ｄＦ（Ｓ$_1$，Ｓ$_2$）を用いる。

【数５】

$$d_F(S_1, S_2) = \sum_{i-1}^{k} w_i d_{Fi}(S_1, S_2)$$

【０１０７】

ここで、｛ｗ$_i$｝は、Σｉｗ$_i$＝１となる重み係数である。

【０１０８】

以上のように、映像音声処理装置は、図５のステップＳ２において抽出された特徴量を用いて非類似性測定基準を計算し、当該セグメント間の類似性を測定することができる。

【０１０９】

次に図５のステップＳ４におけるシーンの切り分けについて説明する。映像音声処理装置は、非類似性測定基準と抽出した特徴量とを用いて、各セグメントに対する近隣の最も類似したセグメントの分布パターンの変化を検出し、シーンの切れ目か否かを判断して出力する。

映像音声処理装置は、シーンを検出する際に、次のような４つの処理を行う。

【０１１０】

①の処理では、各セグメントを基準としたとき、一定の時間枠の中で最も類似したセグメントを一定数検出する。

【０１１１】

②の処理では、①の処理の後、基準セグメントに対し過去と未来の時間帯に存在する類似セグメントの数の比率を計算し(実際には未来に存在している類似セグメントの個数から過去に存在している類似セグメントの個数を減算するなど)、その計算結果を境界性測定値とする。

【０１１２】

③の処理では、②の処理で得られた境界性測定値を、各セグメントを基準としたときの時間変化を調べ、過去比率が高いものがいくつか連続し、未来比率の高いものがいくつか連続するパターンを示すセグメント位置を検出する。

【０１１３】

④の処理では、③の処理のとき、境界性測定値の絶対値を合計し、この合計値をシーン強度値と呼ぶことにする。このシーン強度値があらかじめ決められた閾値を超えた場合、シーンの境界とする。

【０１１４】

これらを順を追って図６を用いて具体的に説明する。①の処理では、図６（Ａ）のように例えば、各セグメントに対して過去に任意のｋ個のセグメント、未来にもｋ個のセグメントの時間枠を設定し(例えばここでは５個)、類似セグメントをこの時間枠の中でＮ個検出する(ここでは４個)。時間は各セグメントを表す数字が大きくなるに連れて未来へと進んで行く。同図の真中の濃い網掛けのセグメント７が、ある時間の基準のセグメントであり、これに対して類似なセグメントはそれよりも薄い網掛けになっているセグメント４，６，９，１０である。ここ

では４個の類似セグメントを抽出しており、過去に２個、未来に２個存在する。

【０１１５】

②の処理では、このとき境界性測定値は、(過去の個数)を(未来の個数)で除したり、(未来の個数)から(過去の個数)を減ずるなどとして計算する。ここでは、後者の方法で境界性測定値を計算する。ここで、各境界性測定値をFiと表す。iは各セグメントの位置(番号)である。いま、後者の方法で計算すると同図(Ａ)の境界性測定値F6は０となる。

【０１１６】

③の処理では、②の処理での計算を時間軸に沿って行って行く。同図（Ｂ）は同図（Ａ）から３セグメント進んだときのセグメント１０に対して過去にセグメント５，８，９の３個、未来にセグメント１１の１個類似セグメントが存在している。このときの境界性測定値 F_{10}＝１－３＝－２となる。

【０１１７】

また、同図（Ｃ）はさらに１セグメント進んでシーンの境界直前に到達した状態であり、セグメント１１の類似セグメント６，７，９，１０はすべて過去に集中している。このとき境界性測定値はF_{11}＝０－４＝－４となる。

【０１１８】

次に、同図（Ｄ）は同図（Ｃ）から１セグメント進んだところであり、境界を越えて新しいシーンに入った直後でシーンの先頭のセグメント１２の場合である。類似セグメントは１３，１４，１５、１６である。このとき類似セグメントは未来にすべて存在するパターンに変化している。F_{12}＝４－０＝４となる。

【０１１９】

最後に、同図（Ｅ）は、さらに１セグメント進んだところのセグメント１３の場合である。同様に、F_{13}＝３－１＝２となる。この方法ではこのように過去の方に類似セグメントの比率が大きいときは負符号（マイナス符号）であり、正符号（プラス符号）は未来に比率が大きいことを示している。このときの境界性測定値F_iの変化は、

０　…　（－２）→（－４）→（＋４）→（＋２）・・・（６）

のようなパターンを示す。

【０１２０】

（－４）→（＋４）と変化しているところがシーンの境界に対応している。これは図６（Ａ）のようにシーンの中間にある場合は時間枠内にある類似的セグメントは各セグメントを挟んで過去、未来にほぼ均等に存在する。しかし、シーンの境界に近づくにつれて同図（Ｂ）のように過去に存在する比率が高くなって行き、同図（Ｃ）で過去の存在比率が１００％になり、同図（Ｄ）のように境界を超えた直後は未来に存在比率が１００％に変わるパターンを持つことを表している。このようなパターンを検出することによりほぼ過去１００％の存在比率から未来への存在比率ほぼ１００％へ大きく変動する変化点がシーンの切れ目と対応付けられる。

【０１２１】

また、シーンの非境界領域の中であっても過去比率が高いパターンから未来比率の高い比率へ一時的に変化(１セグメント間のみ)する場合がある。しかし、それはシーンの境界でないことが多い。なぜならば、このような一時的な変化は偶発的に起こる場合がほとんどだからである。 非境界領域のような類似セグメントが過去に存在比率の大きい境界性測定値が複数続いたあとに、未来に存在比率の大きい境界性測定値が複数続くパターンが検出されたときにシーンの境界の可能性が高いと判断する。そうでないときはシーンの境界ではない可能性が高いため、シーンの境界と見なさい。

【０１２２】

④の処理では、③の処理の後、境界性測定値を合計し、シーン境界点の「強さ」を計算する。その強さを測定するために、境界性測定値の絶対値を足すこととする。その値の変化の度合いがシーンの視覚的変化の度合いに対応しており、また、シーンの視覚的変化の度合いはシーンの意味的な変化の度合いに対応している。したがってこの値の大小によりシーンの意味的度合いの大小に対応したシーンを検出することが可能となる。

【０１２３】

ここではこの絶対値の合計をシーン強度値V_iと定義する。その定義ではｉはセグメントの番号を表す。例えば４つの境界性測定値（各セグメントにおいて過

去の２つのセグメントと未来の１つのセグメントと、そのセグメントの境界性測定値の計４つのセグメントF_{i-2}, F_{i-1}, F_i, F_{i+1}）の絶対値の合計を使っている。

【０１２４】

シーンの境界での境界性測定値の変化のパターンは理論的には、先に示したようにF_{i-1}→F_iの値－４→＋４のように１００％過去に類似セグメントが存在した場合から１００％未来に存在する変化が起こると考えられる。

【０１２５】

このようにシーンの境界では、１セグメント間で大きな変化が起こる。そして式（６）のパターンのように、４セグメント以上に渡って境界性測定値の絶対値が大きいままパターンの変化が起こる可能性は、シーンの境界付近でないと高くない。このパターンの変化の特性から、シーン強度値V_iがある大きさ以上のものだけを実際のシーンの境界と判断することにより、希望とするシーン検出を行うことが出来る。

【０１２６】

図７に実際の音楽番組の３０分程度のビデオデータを使用した結果をグラフ化したものを示す。縦軸にシーン強度値、横軸に各セグメントを表している。色の濃い棒のところのセグメントが実際のシーンの境界(ここではシーンの先頭セグメント)である。この結果の場合、シーン強度値が１２以上をシーンの境界とすると６／７の確率で実際のシーンと一致する。

【０１２７】

以上の流れを図８を使って説明する。ここで説明することは映像音声処理装置で示したシーン検出部１６で行われることであり、この処理はセグメントが生成される毎に以下の処理を行う。

【０１２８】

ステップＳ１１では各セグメントに対し、そのセグメントを中心に±ｋ個のセグメント範囲の中で、特徴量類似性測定部１７を用いて最近傍の類似セグメントをＮ個検出し、それらが過去に存在する個数と未来に存在する個数を求める。

【０１２９】

ステップＳ１２において、ステップＳ１１の処理で求められたＮ個の類似セグメントのうち、未来に存在する個数から過去に存在する個数を減じた数を境界性測定値F_iとして、各セグメントに対し求め、保存する。

【０１３０】

ステップＳ１３では、２ｎ個のセグメントの境界性測定値F_{i-n}，・・・，F_i，F_{i+n}のパターンの変化からシーンの境界の可能性の高い場所を特定する。ｎは、ｉセグメントから過去の比率と未来の比率のパターン変化を見るために必要な境界測定値の数である。

【０１３１】

ここで、シーンの境界を示唆する変化パターンについての３つの条件を次のように定義する。

【０１３２】

①F_{i-n}乃至F_{i+n}がすべて０でないこと

②F_{i-n}乃至F_{i-1}の値が０以下であり、かつ、F_i乃至F_{i+n}の値が０以上であること

③F_{i-n}乃至からF_{i-1}の値が０以下であり、かつ、F_i乃至F_{i+n}の値が０以上であること

【０１３３】

そして、成就値した３つの条件を満足するか否かを判定する。条件を満足した場合、シーンの境界の可能性が高いと判断し、次のステップＳ１４に移行する。そうでない場合は処理がステップ１６に進む。

【０１３４】

ステップＳ１４では、さらにステップＳ１３での境界性測定値を次式に適用して境界性測定値F_{i-n}，・・・，F_i，F_{i+n}からシーン強度V_iを計算する。

$$V_i = |F_{i-n}| + \cdots + |F_{i-1}| + |F_i| + \cdots + |F_{i+n}|$$

【０１３５】

そして、強度値に対する閾値を越える条件が設けられた場合に、その条件を満たすシーン強度値が現れた場合、求めるシーンの視覚的変化の強度であると判断し、処理しているビデオデータのシーンの境界の１つであるとして、その位置を

出力する。強度値に関する条件が必要としない場合に、各セグメントについての強度値を付加情報データとしてビデオデータ記録部１８に出力し記録する。

【０１３６】

以上の処理を繰り返して行くことによりシーンの境界を検出する。シーンはこの境界から境界に含まれるセグメント群がシーンを形成されることとなる。

【０１３７】

以上説明してきたように、本発明を適用した映像音声処理装置は、シーン構造を抽出するためのものである。上述した映像音声処理装置の一連の処理が、テレビドラマや映画など、様々な内容のビデオデータに対して、そのシーン構造を抽出可能であることは、既に実験にて検証済みである。

【０１３８】

なお、本発明は検出されるシーンの境界の数をシーン強度値を任意に変更することにより、調整することが可能である。そのため、このシーン強度値を調整することにより、いろいろなコンテンツにより良く適応したシーンの境界検出を行うことが可能である。

【０１３９】

さらに、ビデオを一覧するに当たって見やすくするために、得られたシーンの数を出来る限り少なくすることが必要となる。それ故に、検出したシーンの数が限定された場合、どのシーンを見せるとよいかという問題が生じる。そのため、得られたシーンの各々の重要性が解れば、その重要性の順番に従い、シーンを見せると一覧するためによい。本技術は得られたシーンがどの程度重要であるかを計る尺度であるシーン強度値を提供してその尺度を変更する(シーン強度閾値を変更する)ことにより、シーンの個数を変更することが可能であり、ユーザの興味に応じて利便性の良い視聴表現を行うことができる。

【０１４０】

しかも、シーンの個数を変更するにあたり、再度シーン検出処理を行うことを必要とせず、シーン強度閾値を変更することのみで保存された強度値時系列を簡単に処理することが可能である。

【０１４１】

以上のように、本発明は、従来技術における上述した全ての問題点を解決したものである。

【０１４２】

まず、映像音声処理装置は、ユーザが事前にビデオデータの意味的な構造を知る必要はない。

【０１４３】

さらに、映像音声処理装置は、各セグメントに対し行われている処理は次の項目を含む。

【０１４４】

①特徴量抽出することと

②一定個数の時間領域内にセグメント対の間の非類似性を測定すること

③非類似性測定結果を用い、一定個数の最も類似したセグメントを抽出すること

④類似したセグメントの存在比率より境界性測定値を計算すること

⑤境界性測定値を用い、シーン境界点の強度値を求めること

【０１４５】

いずれの処理も計算上の負荷が少ない。そのため、セットトップボックスやディジタルビデオレコーダ、ホームサーバ等の家庭用電子機器にも適用することができる。

【０１４６】

また、映像音声処理装置は、シーンを検出した結果、ビデオブラウジングのための新たな高レベルアクセスの基礎を与えることができる。そのため、映像音声処理装置は、セグメントではなくシーンといった高レベルのビデオ構造を用いてビデオデータの内容を視覚化することにより、内容に基づいたビデオデータへの容易なアクセスを可能とする。例えば、映像音声処理装置は、シーンを表示することにより、ユーザは、番組の要旨をすばやく知ることができ、興味のある部分を迅速に見つけることができる。

【０１４７】

さらに、映像音声処理装置は、シーン検出の結果、ビデオデータの概要または

要約を自動的に作成するための基盤が得られる。一般に、一貫した要約を作成するには、ビデオデータからのランダムな断片を組み合わせるのではなく、ビデオデータを、再構成可能な意味を持つ成分に分解することを必要とする。映像音声処理装置により検出されたシーンは、そのような要約を作成するための基礎となる。

【０１４８】

なお、本発明は、上述した実施の形態に限定されるものではなく、例えば、セグメント間の類似性測定のために用いる特徴量等は、上述したもの以外でもよいことは勿論であり、その他、本発明の趣旨を逸脱しない範囲で適宜変更が可能であることはいうまでもない。

【０１４９】

またさらに、本発明はシーン強度値を任意に変更することにより、コンテンツ構造上、重要な変化点であるシーンが得られる。なぜなら、強度値がコンテンツ内容の変化の度合いに対応できるからである。すなわち、ビデオを閲覧する際に、シーン強度値閾値を調整することにより、検出シーンの個数を制御できる。しかも、目的に応じて、コンテンツを表示する個数を増やしたり減らしたりすることが可能となる。

【０１５０】

つまり、コンテンツのいわゆる閲覧粒度(granularity)が目的に応じて自由に制御することができる。例えば、ある一時間ビデオを見るときに、最初に強度値を高く設定し、コンテンツに対して重要であるシーンからなる短い要約を示す。次に、若し興味が増し、詳しく見てみたいと思ったなら、強度値を下げることにより、より細かいシーンによって構成されている要約を表示することができる。しかも本発明の方法を利用すれば、従来技術と異なって、強度値を調整する度に検出を再び行う必要がなく、保存された強度値時系列を簡単に処理を行うことだけ十分である。

【０１５１】

セットトップボックスやデジタルビデオレコーダなどの家庭機器に実装するにあたり、以下のような効果がある。

【０１５２】

１つ目の効果は、本発明のシーン検出は各セグメントに対する類似セグメントの局所的な変化を調べることで実現できるため、調べるセグメントを一定数に固定することができる。そのため処理に必要なメモリ量を固定化することが可能になり、メモリ量の少ないセットトップボックスやデジタルレコーダなどの家庭機器でも実装可能となる。

【０１５３】

２つ目の効果は、１つ目の効果で説明したように、シーンを検出する処理は決められた数のセグメントを処理して行くことによって実現する。これは、各セグメントの処理にかかる時間は一定で行う実時間処理が可能である。これは決められた時間内に決められた処理を必ず終わらせなければならないセットトップボックスやデジタルレコーダなどの家庭機器などに適している。

【０１５４】

３つ目の効果は、上述したようにシーン検出する処理は各セグメント毎に決められた数のセグメントを処理して行くため、１つの処理が終わる毎に新たなセグメントの処理を行う逐次処理が可能である。このことは、セットトップボックスやデジタルレコーダなどの家庭機器において、ビデオ信号などの記録を終了する場合、その終了時刻とほぼ同時に処理を終了することが可能である。また何らかの理由で記録状態が停止した場合でも、それまでの記録を残しておくことが可能である。

【０１５５】

ところで、上述した一連の処理は、ハードウェアにより実行させることもできるが、ソフトウェアにより実行させることもできる。一連の処理をソフトウェアにより実行させる場合には、そのソフトウェアを構成するプログラムが、専用のハードウェアに組み込まれているコンピュータ、または、各種のプログラムをインストールすることで、各種の機能を実行することが可能な、例えば汎用のパーソナルコンピュータなどに、記録媒体からインストールされる。

【０１５６】

この記録媒体は、図３に示すように、コンピュータとは別に、ユーザにプログ

ラムを提供するために配布される、プログラムが記録されている磁気ディスク２２（フロッピディスクを含む）、光ディスク２３（CD-ROM(Compact Disc-Read Only Memory)、DVD(Digital Versatile Disc)を含む）、光磁気ディスク２４（MD(Mini Disc)を含む）、もしくは半導体メモリ２５などよりなるパッケージメディアにより構成されるだけでなく、コンピュータに予め組み込まれた状態でユーザに提供される、プログラムが記録されているROMやハードディスクなどで構成される。

【０１５７】

なお、本明細書において、記録媒体に記録されるプログラムを記述するステップは、記載された順序に従って時系列的に行われる処理はもちろん、必ずしも時系列的に処理されなくとも、並列的あるいは個別に実行される処理をも含むものである。

【０１５８】

また、本明細書において、システムとは、複数の装置により構成される装置全体を表すものである。

【０１５９】

【発明の効果】

以上のように、本発明のＡＶ信号処理装置および方法、並びに記録媒体のプログラムによれば、基準となるセグメントと他のセグメントとの特徴量の類似性を測定するための測定基準を算出し、測定基準を用いて、基準となるセグメントと他のセグメントとの類似性を測定し、測定し類似性を用いて、基準となるセグメントがシーンの境界である可能性を示す測定値を計算するようにしたので、シーンの境界を検出することが可能となる。

【図面の簡単な説明】

【図１】

ビデオデータの階層モデルを示す図である。

【図２】

シーンの境界領域と非境界領域を説明するための図である。

【図３】

本発明の一実施の形態である映像音声処理装置の構成例を示すブロック図である。

【図４】

シーンの境界領域を説明するための図である。

【図５】

映像音声処理装置の動作を説明するフローチャートである。

【図６】

類似セグメントの分布パターンの例を示す図である。

【図７】

シーン検出結果を示す図である。

【図８】

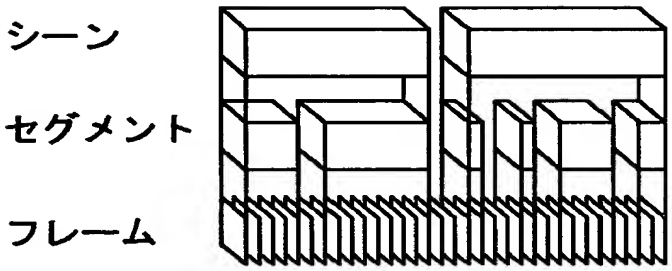
シーン検出部１６の処理を説明するフローチャートである。

【符号の説明】

１１　ビデオ分割部，　１２　ビデオセグメントメモリ，　１３　映像特徴量抽出部，　１４　音声特徴量抽出部，　１５　セグメント特徴量メモリ，　１６　シーン検出部，　１７　特徴量類似性測定部，　１８　ビデオデータ記録部，　１９　ビデオ表示部，　２０　制御部，　２１　ドライバ，　２２　磁気ディスク，　２３　光ディスク，　２４　光磁気ディスク，　２５　半導体メモリ

【書類名】図面

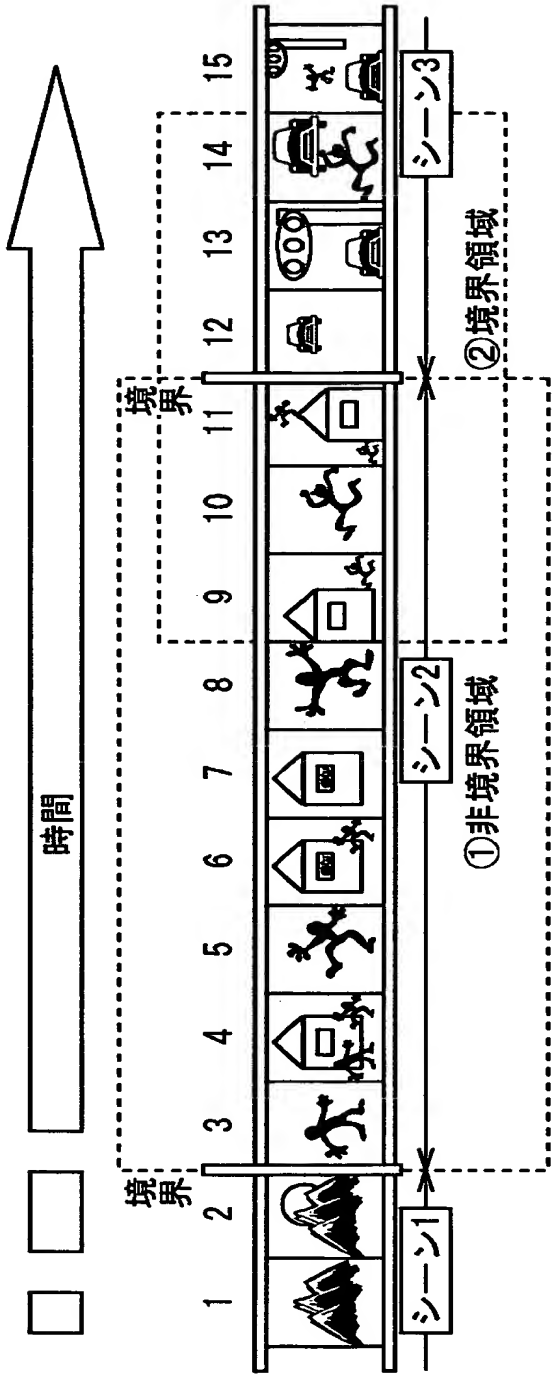
【図１】



ビデオデータの階層モデル

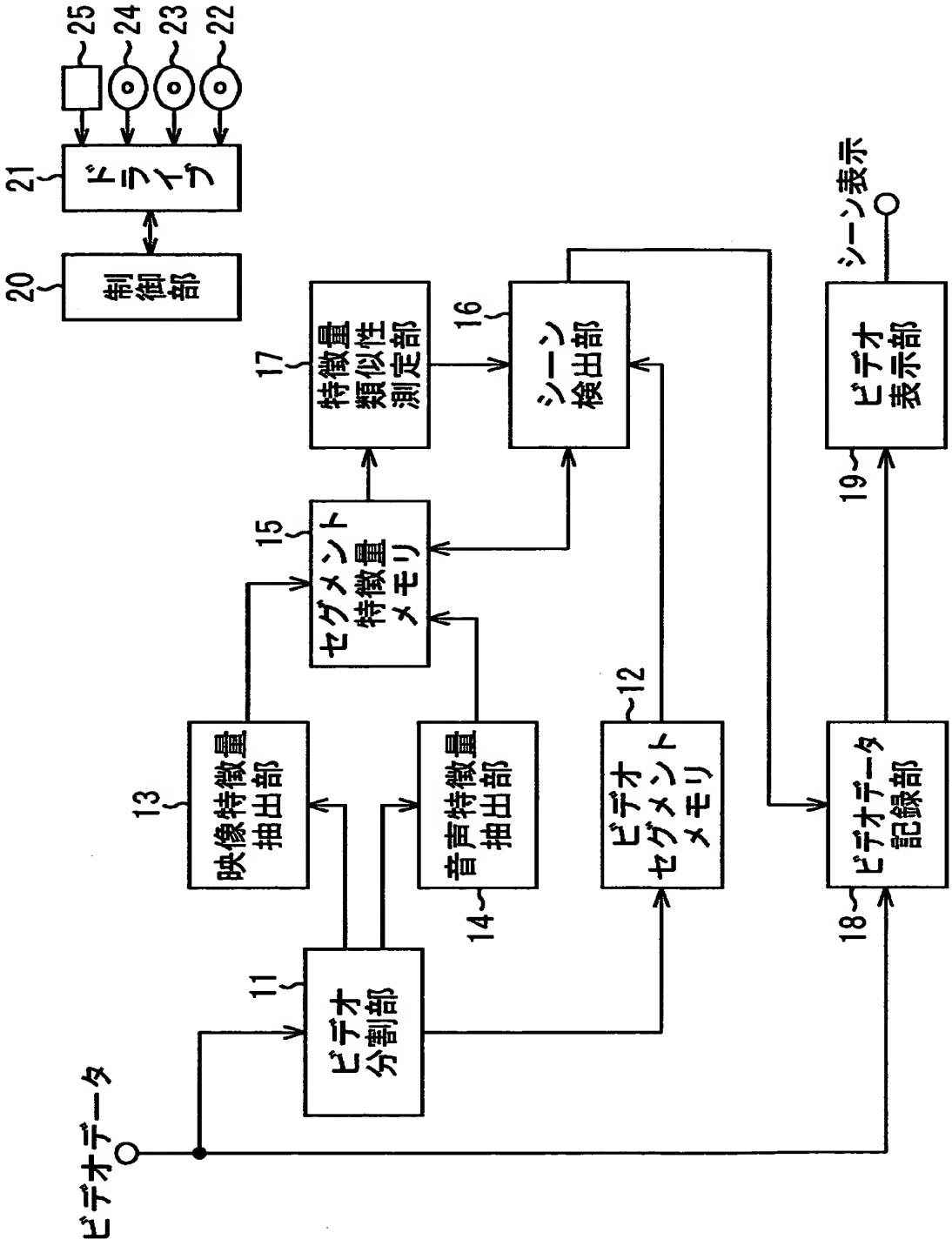
出証特２００１－３０４２４５５

【図２】

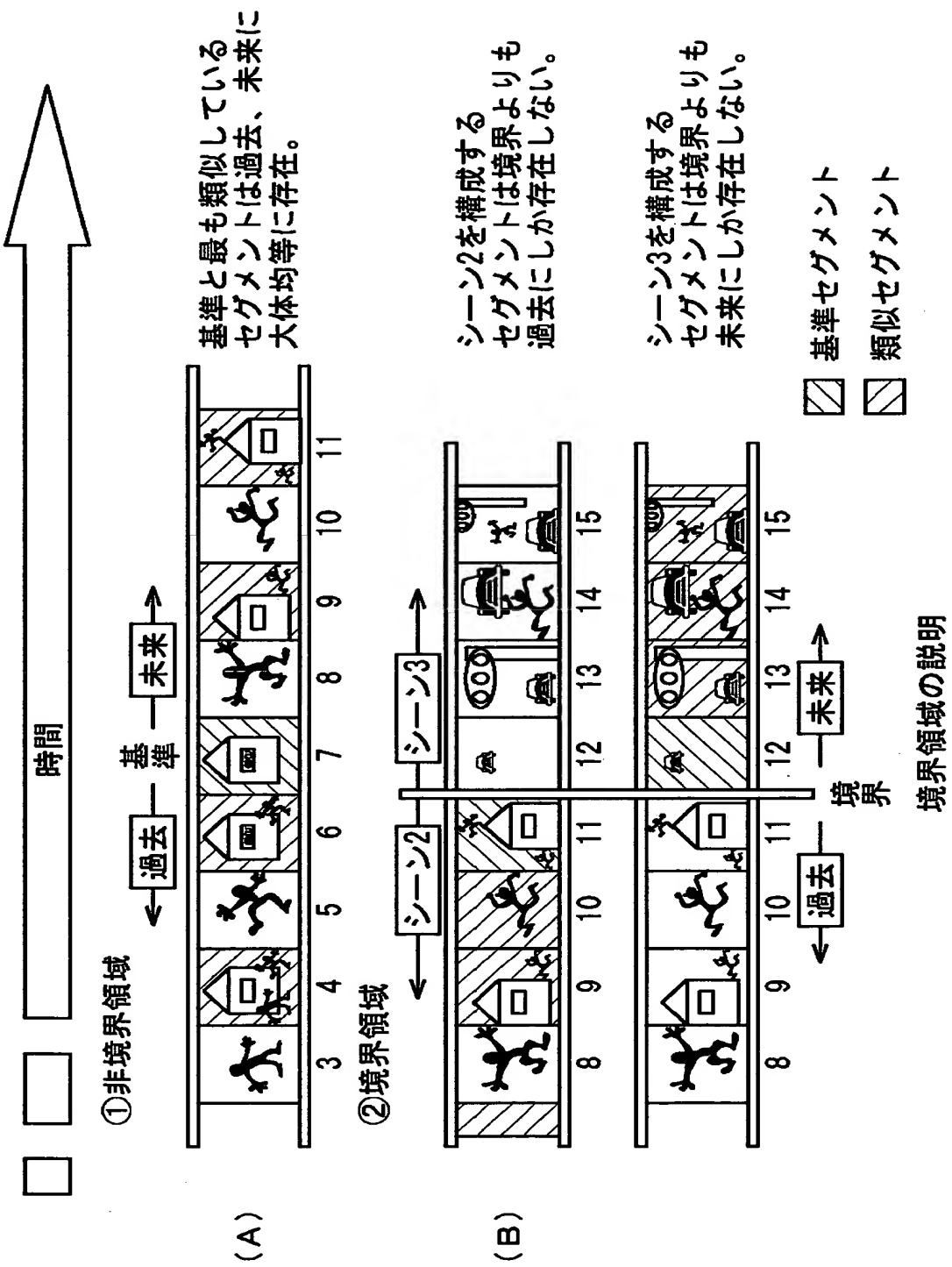


シーンの境界領域と非境界領域

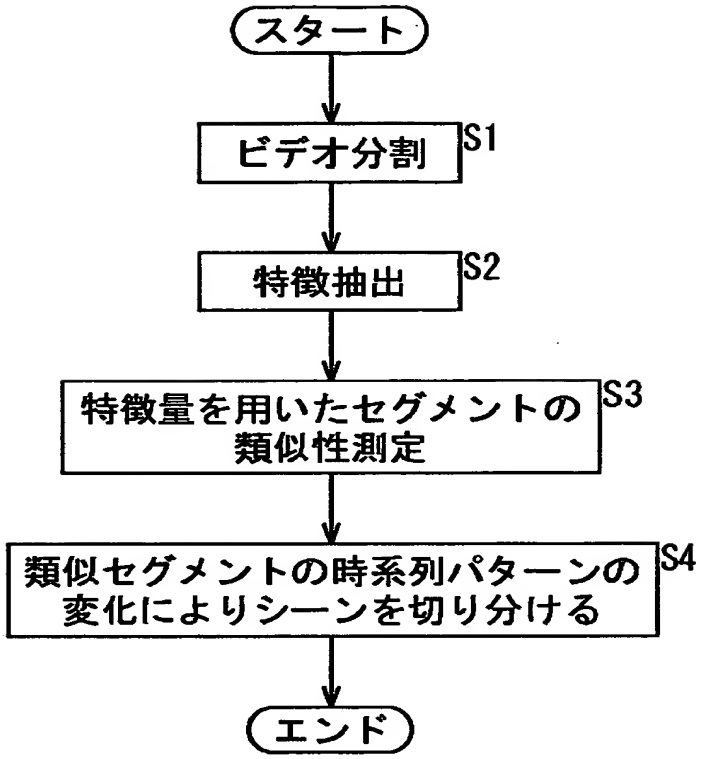
【図３】


25
24
23
22
21
ドライブ
20
制御部
ビデオデータ
11
ビデオ分割部
13
映像特徴量抽出部
14
音声特徴量抽出部
15
セグメント特徴量メモリ
17
特徴量類似性測定部
16
シーン検出部
12
ビデオセグメントメモリ
18
ビデオデータ記録部
19
ビデオ表示部
シーン表示

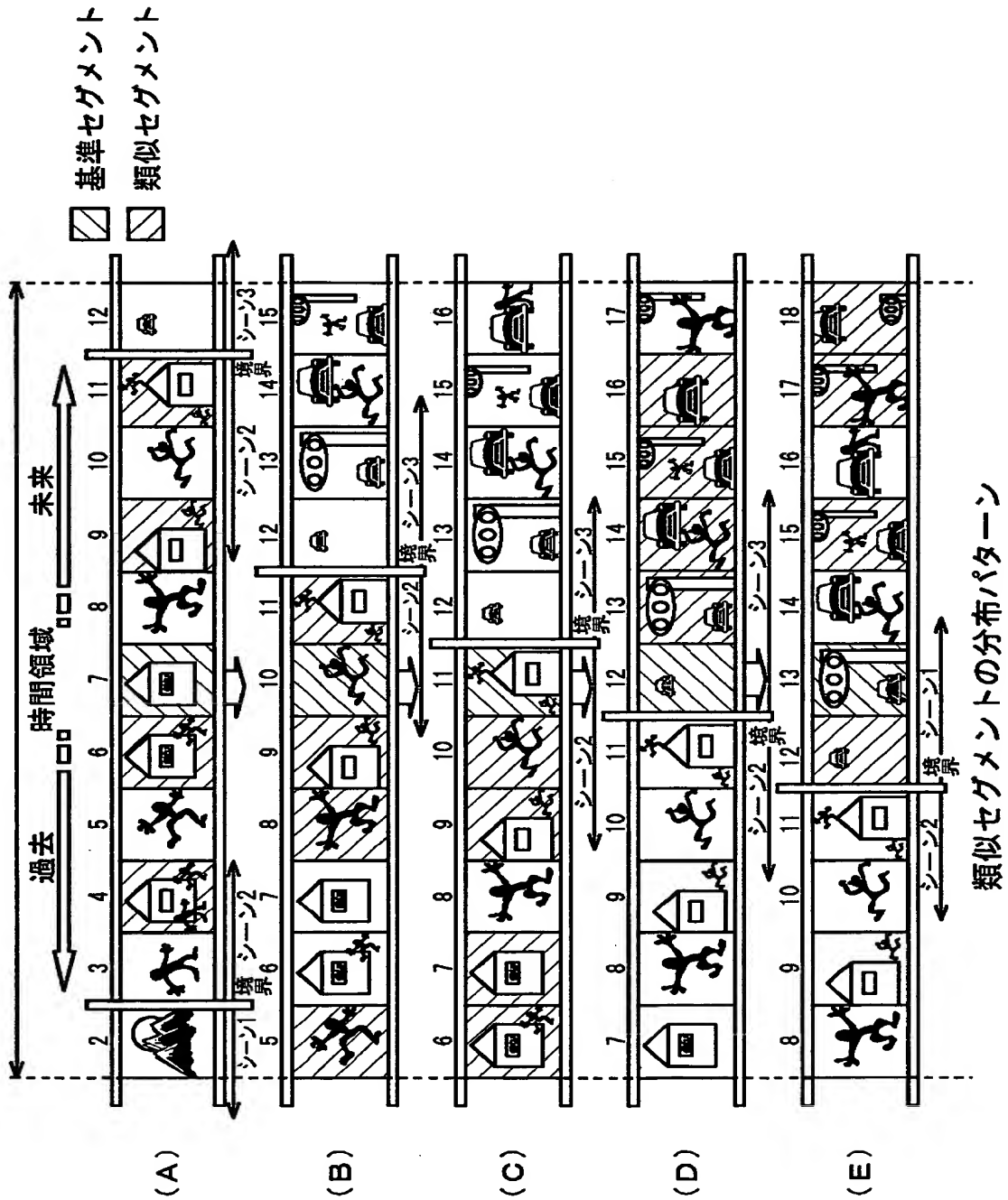
【図４】


時間
①非境界領域
過去
基準
未来
基準と最も類似している
セグメントは過去、未来に
大体均等に存在。
3
4
5
6
7
8
9
10
11
(A)
②境界領域
シーン2
シーン3
シーン2を構成する
セグメントは境界よりも
過去にしか存在しない。
シーン3を構成する
セグメントは境界よりも
未来にしか存在しない。
8
9
10
11
12
13
14
15
境界
過去
未来
(B)
基準セグメント
類似セグメント
境界領域の説明

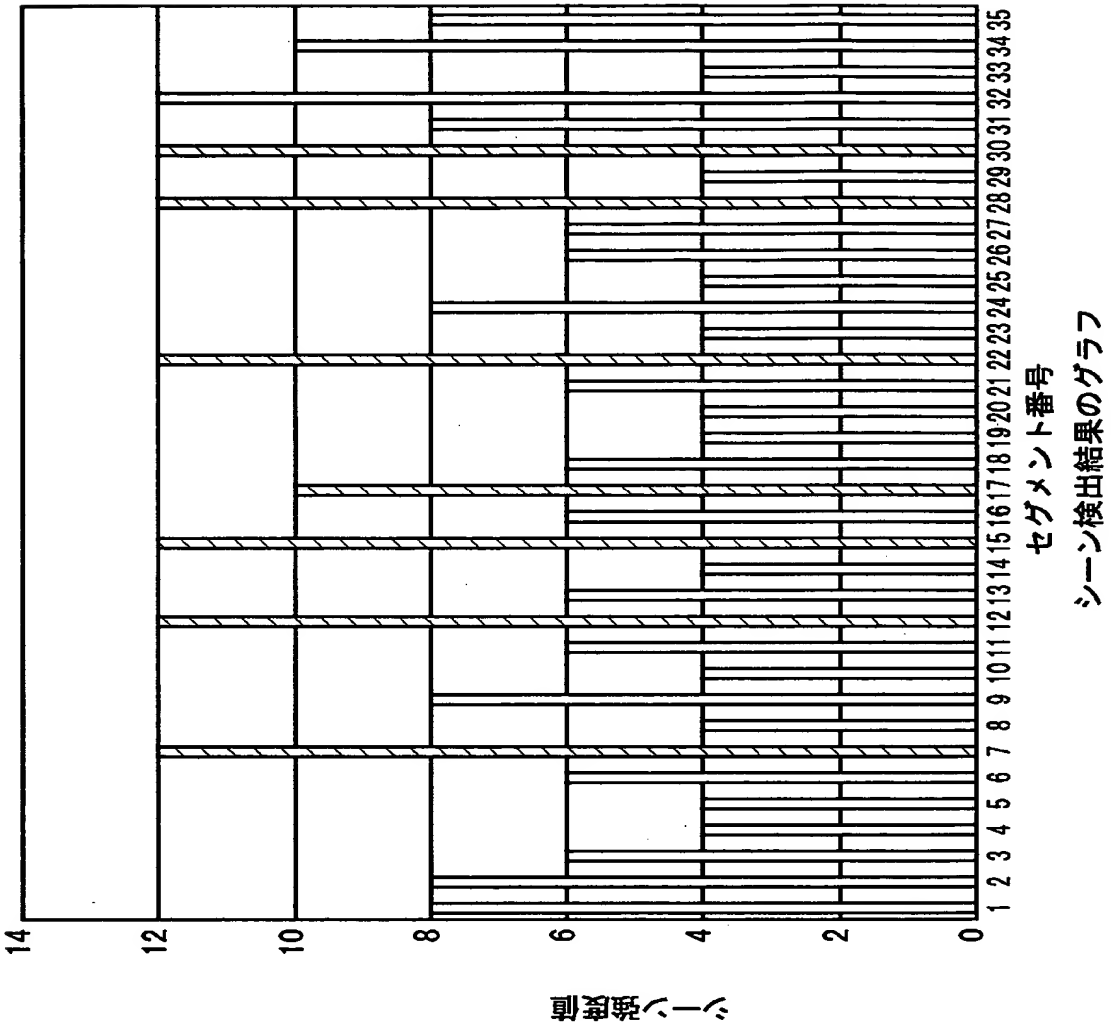
【図５】


スタート
ビデオ分割
S1
特徴抽出
S2
特徴量を用いたセグメントの類似性測定
S3
類似セグメントの時系列パターンの変化によりシーンを切り分ける
S4
エンド

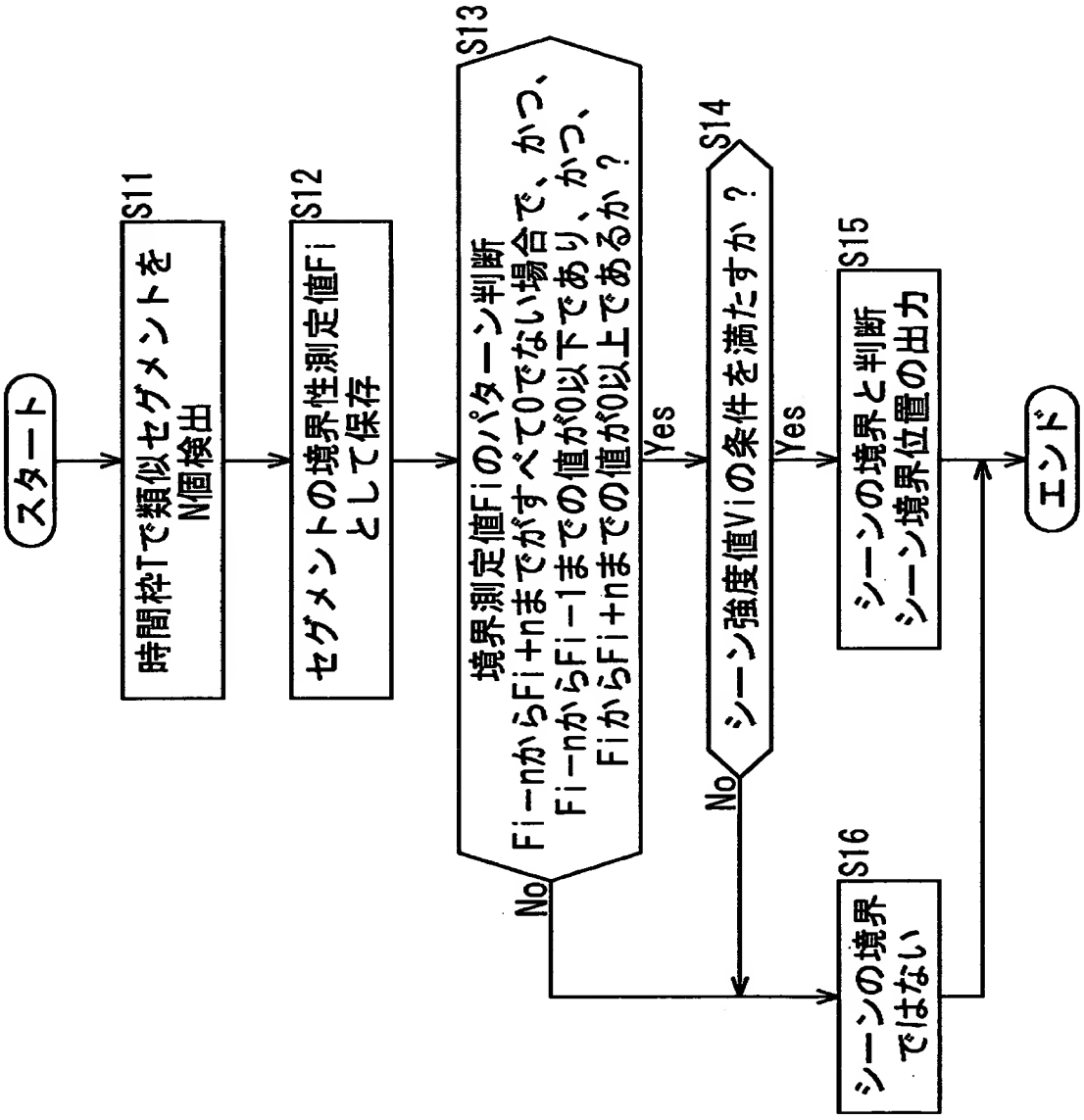
【図６】


基準セグメント
類似セグメント
未来
時間領域
過去
シーン1
シーン2
シーン3
境界
類似セグメントの分布パターン
(A)
(B)
(C)
(D)
(E)

【図７】


シーン強度値
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35
セグメント番号
シーン検出結果のグラフ

【図８】


スタート
S11
時間枠Tで類似セグメントをN個検出
S12
セグメントの境界性測定値Fiとして保存
S13
境界測定値Fiのパターン判断
Fi−nからFi+nまでがすべて0でない場合で、かつ、
Fi−nからFi−1までの値が0以下であり、かつ、
FiからFi+nまでの値が0以上であるか？
Yes
No
S14
シーン強度値Viの条件を満たすか？
Yes
No
S15
シーンの境界と判断
シーン境界位置の出力
S16
シーンの境界ではない
エンド

【書類名】　　要約書

【要約】

【課題】　シーンの境界を検出する。

【解決手段】　ステップＳ１において、入力されたビデオデータを映像セグメントまたは音声セグメントのいずれか、あるいは可能であればその両方に分割する。ステップＳ２において、セグメントの特徴を表す特徴量を計算する。ステップＳ３において、特徴量を用いたセグメントの類似性測定を行う。ステップＳ４において、セグメントがシーンの切れ目にあたるか否かを判断する。すなわち、映像音声処理装置は、先のステップＳ３において計算した非類似性測定基準と、先のステップＳ２において抽出した特徴量とを用いて、各セグメントを現在と見なし、近接の類似したセグメントが、その基準とするセグメントに対し過去か未来かどちらに存在比率が高いかを求め、その存在比の率変化のパターンを調べ、シーンの境界であるか否かの判断をする。

【選択図】　　図５

出願人履歴情報

識別番号　〔０００００２１８５〕

１．変更年月日　１９９０年　８月３０日
［変更理由］　新規登録
住　所　東京都品川区北品川６丁目７番３５号
氏　名　ソニー株式会社